Topic: Insurance Claim Fraud Detection: Machine Learning Project

By: Shashank Vinod Chardeve And Chinmay H R

Introduction:

1) Every year, insurance fraud results in large financial losses and increased premiums for law-abiding policyholders.
2) Manual detection techniques are laborious and frequently unable to handle the increasing complexity of data in insurance claims.
3) Since most fraudulent claims go unnoticed, automated techniques present potential for improvement.
4) Large-scale claims data is analyzed by machine learning to swiftly identify questionable activities and new fraud trends.
5) Insurers can reduce payouts to fraudsters and expedite the processing of valid claims by implementing automation for early identification.
6) This project develops a solid, scalable solution for accurate insurance fraud detection using cutting-edge machine learning algorithms.

Objective:

- The purpose of this task is to provide you with a hands-on understanding of model selection in realistic situations.
- In order to find patterns and insights, you will learn how to preprocess and analyze data, extract significant features, and use a variety of machine learning models.
- You will be able to evaluate many strategies and choose the best one by using model assessment and optimization techniques.
- Your ability to create trustworthy prediction models, make better decisions, and guarantee model correctness and generalization in a variety of applications will all be improved by this procedure.

Data Preparation & Cleaning:

Removed the identification and superfluous columns.

Date types, null values, and nonsensical data were fixed.

Stratified 70-30 split for validation and training.

Exploratory Data Analysis (EDA) :

- Univariate analysis: Histograms & stats for all key features.
- Bivariate analysis: Boxplots and likelihood tables for categorical variables.
- Correlation analysis: Heatmap to spot related features.
- Class imbalance: 75% non-fraud, 25% fraud.

Feature Engineering :

- Random Over Sampler was used to resample and balance the fraud class.
- New features were created (for example, the number of days between policy bind and incident).
- Consolidated unusual categories and used fake encoding.
- Scaled numerical characteristics.

Model Building – Logistic Regression :

- RFECV was used to pick features, followed by p-values and VIF checks.
- Initial cutoff (0.5): sensitivity = 0.42, specificity = 0.95, and F1 = 0.57.
- Optimised cutoff (0.2): sensitivity = 0.94, specificity = 0.78, F1 = 0.87, accuracy = 86%.

Model Building – Random Forest :

- Claim amounts, incident timing, insurance duration, and so on are all shown as important features.
- Cross-validation accuracy is 92%.
- Hyperparameter adjustment resulted in maximum accuracy.
- Training F1 score: 1.00 (validation F1: 0.62).

How can we analyze historical claim data to detect patterns that indicate fraudulent claims?

- Collect and aggregate detailed claim records, including policyholder information, claim history, incident details, and payout amounts.
- Clean and preprocess data by handling missing values, transforming dates and categorical fields, and removing outliers and inconsistencies.
- Perform exploratory data analysis (EDA) to identify unusual patterns, correlations, and trends in claims—such as spikes in claim amounts, early incidents after policy start, or repeated suspicious claim types.
- Use statistical tests and visualizations (histograms, boxplots, heatmaps) to detect anomalies that might warrant further investigation.

- To improve fraud pattern identification, create new information including ratios and discrepancies between claim dates, amounts claimed versus policy limitations, and policy duration at claim time.

- Utilize machine learning techniques (such as clustering, anomaly detection, and classification) to automatically identify claims that exhibit patterns resembling known fraudulent cases by learning from past data.

- To find hidden relationships and discrepancies, apply text mining and association rule mining to claim narratives and supporting evidence.

- Refine detection accuracy by using newly discovered fraudulent claims to continuously improve models and analysis methods as fraud strategies change.

Which features are most predictive of fraudulent behavior?

- Total Claim Amount: Fraud is frequently indicated by exceptionally large or unusual claim amounts in comparison to policy limits.
- Policy incident duration: Suspicious claims are typically submitted fairly soon after the policy's binding date.
- Incident Severity: Allegations of insignificant occurrences or small damage could be signs of deceptive exaggeration.
- Collision Type: According to data assessments, some accident types—such as side or rear collisions—have a higher probability of fraud.
- Policy Deductible and limits: Because of coverage manipulation, lower deductibles or umbrella limitations may be associated with a rise in false claims.

- Insured Occupation and Hobbies: Certain professions (executive-managerial, for example) and pastimes (chess, cross-fit) exhibit statistical correlations with fraud in patterns identified by machine learning algorithms.

- Auto Make and Model: A higher frequency of false claims may indicate specialized fraud schemes involving specific car brands and models.

- Incident Time and Location: Patterns can be found by analyzing the time of day and the geographic location of incidents; claims from certain regions or during odd hours may be flagged.

Can we predict the likelihood of fraud for an incoming claim, based on past data?

- Using past claim data, machine learning models can be taught to find trends that are highly connected to fraud.
- To assign a fraud probability to new claims, these models understand intricate correlations between incident facts, policy specifics, consumer behavior, and claim elements.
- Based on these discovered patterns, supervised learning methods (such as logistic regression, random forests, and XGBoost) categorize assertions as either authentic or false.
- Several measures (accuracy, recall, precision, and F1-score) are used to verify and optimize models in order to balance false alarms and detection capability.

- The probability score, or anticipated likelihood, makes it possible to rank high-risk claims for manual review or more research.
- Workflows for processing claims can use models to identify dubious instances early, minimizing losses and increasing productivity.
- Constantly retraining the model with fresh data keeps detection effective over time by adapting it to new fraud schemes.
- Such models have shown notable savings in operational expenses and fraudulent payouts in real-world implementations.

What insights can be drawn from the model that can help in improving the fraud detection process?

- Early Identification: By detecting questionable claims early in the approval process, the methodology lowers financial risk.
- High-Risk Profiles: Provides focused surveillance by identifying particular client profiles, professions, and behaviors that are more likely to be fraudulent.
- Feature Importance: Identifies the most predictive claim features (such as event severity and claim amount), directing targeted data gathering and validation.
- Optimized Thresholds: Modifies categorization cutoffs according to business priorities (e.g., risk tolerance) in order to balance false positives and false negatives.

- Resource Allocation: Prioritizes high-probability cases identified by the model to effectively allocate resources for fraud investigations.

- Dynamic Adaptation: The detection system can adapt to new fraud strategies and developing patterns through ongoing model retraining.

- Data Quality Focus: Promotes enhancements in data collection and preprocessing by highlighting significant features and areas with missing data.

- Cross-Departmental Cooperation: To improve controls, insights encourage cooperation between data teams, claims adjusters, and legal departments.

Conclusion:

- Automated ML models can significantly reduce fraudulent payouts.
- Strong validation metrics support deployment.
- Ongoing monitoring and tuning are advised.

References:

- Automated ML models can significantly reduce fraudulent payouts.
- Strong validation metrics support deployment.
- Ongoing monitoring and tuning are advised.